5-О

УДК 004.93

А.О. Олійник, Є.О. Гофман, С.О. Субботін
Запорізький національний технічний університет, Україна
Україна, 69063, м. Запоріжжя, вул. Жуковського, 64, *olejnikaa@gmail.com*

Метод кластеризації даних на основі дерев розв’язків

A.A. Oliinyk, Ye.A. Gofman, S.A. Subbotin
Zaporizhzhya National Technical University
Ukraine, 67063 Zaporizhzhya, Zhukovskogo st., 64

Clustering Method Based on Decision Trees

А.А. Олейник, Е.А. Гофман, С.А. Субботин
Запорожский национальный технический университет, Украина
Украина, 69063, г. Запорожье, ул. Жуковского, 64

Метод кластеризации данных на основе деревьев решений

Досліджено застосування дерев розв’язків для розв’язання завдання кластерного аналізу. Розроблено метод кластерного аналізу, що дозволяє виконувати розбиття простору екземплярів на кластери, при використанні якого відсутня необхідність задання інформації про кількість кластерів та їх форму, що суттєво розширює можливість його застосування на практиці. Проведено експерименти з розв’язання завдань кластер-аналізу з використанням запропонованого методу.
Ключові слова: дерево розв’язків, кластеризація, навчальна вибірка, обробка даних.

The usage of decision trees for the problem of cluster analysis is investigated. The method of cluster analysis that allows the partition of instances into clusters, using which there is no need to specify information about the number of clusters and their shape that significantly expands possibilities of its usage in practice, is developed. The experiments for solving the cluster analysis problems using the proposed method are made.
Key words: decision tree, clustering, learning sample, data processing.

Исследовано применение деревьев решений для задачи кластерного анализа. Разработан метод кластерного анализа, позволяющий выполнять разбиение пространства экземпляров на кластеры, при использовании которого отсутствует необходимость задания информации о количестве кластеров и их форме, что существенно расширяет возможности его применения на практике. Проведены эксперименты по решению задач кластер-анализа с использованием предложенного метода.
Ключевые слова: дерево решений, кластеризация, обучающая выборка, обработка данных

Вступ

При розв’язанні завдань технічного діагностування, розпізнавання образів та прогнозування актуальною є задача кластерного аналізу, що полягає в розбитті деякої вибірки даних на множину кластерів, які являють собою компактні області (таксони) в просторі ознак. Відомі різні методи кластерного аналізу [1], [2], основним недоліком яких є необхідність попереднього задання вхідних параметрів, що настроюються (наприклад, кількість кластерів, які повинні бути виділені). Це ускладнює їхнє застосування при обробці даних у реальних ситуаціях, коли немає достатньої інформації про досліджуваний об’єкт, процес або систему.

Тому актуальною є розробка нових методів кластеризації, вільних від зазначених недоліків, що забезпечують необхідну точність одержуваних розв’язків. Основним критерієм, якому повинні задовольняти методи, застосовувані для розв’язання даного завдання, є можливість поділу простору екземплярів на області з подібними характеристиками. До таких методів відноситься пошук на основі дерев розв’язків, які за рахунок своєї структури виконують розбиття простору рішень на області залежно від значень вхідних змінних [3-5]. У зв’язку із цим у даній роботі пропонується розв’язувати завдання кластерного аналізу на основі побудови дерев розв’язків.

Відомі різні методи ідентифікації дерев розв’язків (ID3, CART, CHAID, QUEST, C5.0). Однак вони не враховують особливостей розв’язуваного завдання кластерного аналізу, пов’язаного з виділенням таксонів, що складаються з об’єктів з найбільш подібними характеристиками [3-7].

Метою даної роботи є розробка методу кластерного аналізу, заснованого на побудові дерев розв’язків, який дозволить виконувати розбиття на кластери шляхом введення рівномірно розподілених точок простору пошуку та дозволить скоротити вимоги до обчислювальних ресурсів при виконанні кластерного аналізу.

Для досягнення поставленої мети необхідно розв’язати такі завдання:

– огляд існуючих методів кластерного аналізу та виявлення їх переваг і недоліків;

– вивчення основних понять, принципів і особливостей дерев розв’язків;

– модифікація розглянутого методу відповідно до специфіки розв’язуваного завдання;

– порівняння розробленого підходу з існуючими методами кластерного аналізу шляхом проведення експериментів і аналізу отриманих результатів.

Постановка задачі. Нехай задана множина об’єктів O, кожний з яких характеризується множиною значень ознак X. Тоді завдання кластерного аналізу полягає в тому, щоб на основі значень ознак X розбити множину об’єктів O на m кластерів (підмножин) $C_1, C_2, \ldots, C_m$, так, щоб кожний об’єкт O_i належав одній і тільки одній підмножині розбиття і щоб об’єкти, які належать одному кластеру, були подібними, у той час, як об’єкти, що належать різним кластерам, були різнорідними.

1 Кластерний аналіз

Кластерний аналіз полягає в розбитті даних на групи схожих об’єктів. Кожна група, що називається кластером, складається з об’єктів, які схожі між собою і які при цьому різні з об’єктами інших груп.

Існує декілька видів методів кластерного аналізу, що відрізняються між собою допущеннями про форму кластерів, видом результуючого розбиття та параметрами, які повинні бути встановлені (наприклад, кількістю кластерів).

Виключаюча кластеризація: дані групуються шляхом виключення одиниць даних. Якщо певний об’єкт належить одному кластеру, то він не може бути включений в інший кластер (до таких методів відноситься, наприклад, метод k-середніх). Основними недоліками такого підходу є:

– необхідність задання кількості кластерів, на які необхідно розбити вхідну вибірку;

– виконується пошук кластерів тільки заданої форми.

Перекриваюча кластеризація: дані можуть входити у два або більш кластерів залежно від значення функції приналежності. До таких методів відноситься метод нечітких c-середніх. При використанні цих методів також необхідно задавати кількість кластерів.

Ієрархічна кластеризація: на початку кластеризації кожний об'єкт розглядається як окремий кластер, після чого два найближчі кластери поєднуються в один і так далі. Метод закінчує свою роботу, коли всі дані об'єднані в один кластер або якщо виконалася умова закінчення роботи. Основним недоліком такого підходу є істотна обчислювальна складність, що особливо помітно при обробці багатовимірних вибірок великого обсягу.

Імовірнісна кластеризація має два різновиди:

– методи, засновані на суміші багатовимірних нормальних розподілів;

– методи інтелектуальної оптимізації, засновані на моделюванні колективного інтелекту суспільних живих істот.

Оскільки даний підхід заснований на імовірнісному підході, то існує можливість незбіжності до оптимального розв'язку.

Як видно з наведеної класифікації, кожний з розглянутих методів має певні недоліки, основні з яких є: необхідність задання кількості формованих кластерів, допущення про форму кластерів, велика обчислювальна складність. У зв'язку з цим можна зробити висновок про те, що застосування дерев розв'язків для кластерного аналізу є перспективним, однак дану техніку необхідно застосовувати з урахуванням особливостей розв'язуваного завдання кластерного аналізу.

2 Дерева розв'язків

Дерева розв'язків являють собою спадну систему, засновану на підході «розділяй і пануй», основною метою якої є поділ дерева на взаємно неперетинні підмножини [3], [5]. Кожна підмножина являє собою підзадачу класифікації.

Дерево розв'язків описує процедуру ухвалення рішення про приналежність певного екземпляра до того або іншого класу.

Дерево розв'язків є деревоподібною структурою, що складається з внутрішніх і зовнішніх вузлів, зв'язаних ребрами [6]. Внутрішні вузли – модулі, що приймають рішення, – розраховують значення функції розв'язку, на підставі чого визначають дочірній вузол, який буде відвіданий далі. Зовнішні вузли (листи), навпаки, не мають дочірніх вузлів і описують або мітку класу, або значення, що характеризує вхідні дані.

У загальному випадку дерева розв'язків використовуються в такий спосіб. Спочатку передаються дані (звичайно це вектор значень вхідних змінних) на кореневий вузол дерева розв'язків. Залежно від отриманого значення функції розв'язку, використовуваної у внутрішньому вузлі, відбувається перехід до одного з дочірніх вузлів. Такі переходи тривають доти, поки не буде відвіданий кінцевий вузол, що описує або мітку класу, або значення, зв'язане із вхідним вектором значень ознак.

3 Кластеризація на основі побудови дерев розв'язків

У пропонованому методі кластеризації даних на основі побудов дерев розв'язків у процесі синтезу дерев використовується традиційний підхід, що дозволяє розділити простір пошуку на кілька різних класів на основі функції пріоритетності. Однак, оскільки при розв'язанні завдання кластеризації не задані класи екземплярів, то пропонується вводити неіснуючі рівномірно розподілені екземпляри для проведення кластерного аналізу. За рахунок введення таких екземплярів можна умовно розбити вхідну вибірку, як мінімум, на два класи: існуючі екземпляри й неіснуючі екземпляри, за рахунок чого можна виконувати класифікацію з використанням дерев розв'язків. При цьому такий підхід дозволяє виділити ті області, які являють собою кластери, оскільки в цих областях більше перебуває реальних екземплярів, ніж штучно доданих. Далі представлені основні особливості пропонованого методу.

При побудові дерева розв'язків для кожної ознаки з n-вимірного простору (n – кількість ознак, що характеризують навчальну вибірку) метод розраховує індекс Джині для розбиття дерева розв'язків, використовуваний як критерій пріоритетності альтернативних можливих варіантів розбиття за ознакою. Тобто поточний вузол дерева розбивається за ознакою, за якою отримано краще (найменше) значення індексу Джині.

У кожному вузлі відбувається розбиття за певною ознакою на ліву й праву гілки (області обмежені попередніми розбиттями). Таким чином, даний етап припускає виконання наступної послідовності дій:

– установити лічильник ознак в одиницю: $i = 1$;

– для кожного конкретного значення ознаки X_i розраховується індекс Джині;

– установити: $i = i + 1$;

– якщо $i \le n$, то виконати перехід до розрахунків індексу Джині для наступної ознаки;

– зберегти краще розбиття для поточного вузла;

– виконати розбиття для лівого нащадка;

– виконати розбиття для правого нащадка.

Принциповою особливістю етапу обчислення індексу Джині є те, що індекс Джині для розбиття обчислюється для значень ознак і деяких рівномірно розподілених штучно доданих K точок. Кожне значення ознаки X_i розглядається як можливе розбиття, тому індекс Джині розраховується для кожного значення.

Нехай є множина M з відповідною потужністю $|M|$. Нехай додатково до цієї множини додається множина K рівномірно розподілених точок потужності $|K| = |M|$ (кількість додаткових точок успадковується від батьківського вузла). Кожне значення $x \in M$ розбиває множину на дві області.

Нехай у лівій області відносно поточної точки $x \in M$ знаходяться області з k_{x-} та m_{x-} точок, значення яких менше заданого значення, у правій області знаходяться $m_{x+} = |M| - m_{x-}$ та $k_{x+} = |K| - k_{x-}$ точок відповідно. Тоді розрахувати k_{x+} та k_{x-} можна таким чином:

$$k_{x-} = |M| - k_{x+} = \frac{|K|(x - \min(M))}{\max(M) - \min(M)},$$

$$k_{x+} = |M| - k_{x-} = \frac{|K|(\max(M) - x)}{\max(M) - \min(M)},$$

де x – конкретне значення ознаки; $\min(M)$ – мінімальне значення з M, $\max(M)$ – максимальне значення в M. Така формула означає, що якщо в межах між $\min(M)$ і $\max(M)$ знаходиться $|k|$ рівномірно розподілених точок, тоді в інтервалі між $\min(M)$ і поточним значенням x знаходиться n_{x-} точок.

У загальному випадку індекс Джині для розбиття за x можна розрахувати за формулою:

$$g_x = \frac{k_{x-} + m_{x-}}{|K| + |M|} g_{x-} + \frac{k_{x+} + m_{x+}}{|K| + |M|} g_{x+},$$

де індекси Джині для підмножин $x-$ і $x+$ розраховуються в такий спосіб:

$$g_{x*} = 1 - \frac{k_{x*}^2 + m_{x*}^2}{(k_{x*} + m_{x*})^2},$$

де * позначає відповідну підмножину (+ або –).

Після того як отримано краще розбиття, воно переноситься в поточний вузол, його правий і лівий нащадки успадковують множину *K*, що включає n_{x-} та n_{x+} точок відповідно.

Обчислення розбиттів триває доти, поки:

– поточний вузол містить екземпляри в кількості $|M|$, більшій за задане мінімальне значення. Тобто даний параметр є єдиним вхідним параметром, що настроюється, для пропонованого методу;

– поточна множина даних містить групи з як мінімум двома точками (при цьому точки з однаковими значеннями групуються на початковому етапі).

Таким чином, розбиття вузла повинно тривати при виконанні хоча б однієї з даних умов. А якщо ні, то розбиття на даній гілці повинне завершитися.

Виходячи з вищесказаного, можна відзначити, що основною особливістю запропонованого методу є введення додаткових рівномірно розподілених екземплярів, що дозволяє виконувати класифікацію, як мінімум, для двох класів екземплярів. При цьому основною перевагою запропонованого методу є те, що немає необхідності задання інформації про кількість кластерів, їх форму та ін., що суттєво розширює можливість застосування розробленого методу кластерного аналізу на основі побудови дерев розв'язків.

4 Експерименти та результати

Запропонований метод кластерного аналізу на основі побудови дерев розв'язків був програмно реалізований у середовищі пакета Matlab 7.0.

За допомогою розробленого програмного забезпечення і вбудованих засобів пакета Matlab 7.0 проводилися експерименти, які полягали в розбивці на кластери штучно сформованих вибірок за допомогою розробленого методу, а також за допомогою методів кластеризації: k -середніх і агломеративного ієрархічного методу.

Вибірки формувалися випадковим чином на основі нормального розподілу з різними математичними очікуваннями та дисперсіями. Було сформовано чотири двовимірні вибірки, що відрізняються між собою ступенем перетину кластерів. Параметри розподілів, на підставі яких формувалися вибірки, наведено в табл. 1. Кожна вибірка складалася з чотирьох кластерів, кожний з яких, у свою чергу, складався з 200 екземплярів, що характеризуються двома ознаками. Як можна бачити з табл. 1, друга й четверта вибірки характеризуються більшим перетином кластерів порівняно з першою та третьою вибірками.

Таблиця 1 – Параметри розподілів вибірок

Вибірка	Кластер	x_1		x_2		Вибірка	Кластер	x_1		x_2	
		M(X)	D(X)	M(X)	D(X)			M(X)	D(X)	M(X)	D(X)
1	1	0	3	0	3	3	1	0	3	0	3
	2	15	3	15	3		2	0	3	25	4
	3	15	3	0	3		3	16	3	25	4
	4	0	3	15	3		4	25	5	0	3
2	1	0	3	0	3	4	1	0	3	0	3
	2	13	3	13	3		2	0	3	12	3
	3	13	3	0	3		3	12	3	0	3
	4	0	3	13	3		4	12	3	12	3

Як критерій порівняння результатів роботи досліджуваних методів кластеризації використовувалася помилка класифікації:

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N} res_i ,$$

де $res_i = 1$, якщо $cluster_i^* \neq cluster_i$, в іншому випадку – $res_i = 0$; $cluster_i^*$ – номер кластера, до якого віднесений *i*-й об'єкт за допомогою заданого методу кластерного аналізу, $cluster_i$ – номер кластера, до якого відноситься *i*-й об'єкт у заданій навчальній вибірці.

Результати роботи відомих методів кластеризації та запропонованого методу представлені в табл. 2.

Таблиця 2 – Результати роботи методів кластерного аналізу

Метод	Значення помилки			
	Вибірка 1	Вибірка 2	Вибірка 3	Вибірка 4
Метод k-середніх	0,0113	0,0288	0,0050	0,0288
Ієрархічний агломеративний метод	0,0138	0,0325	0,0075	0,0300
Метод кластерного аналізу на основі дерев розв'язків	0,0043	0,0215	0,0041	0,0219

Виходячи з результатів експериментів, представлених у табл. 2, можна бачити, що запропонований метод характеризується меншою помилкою класифікації порівняно з методами: k-середніх та ієрархічним агломеративним. При цьому найбільша помилка класифікації спостерігалася для всіх методів при аналізі другої та четвертої вибірок, для яких характерне суттєве перетинання кластерів.

Також важливо відзначити, що для роботи запропонованого методу не треба було задавати кількість вихідних кластерів, на відміну від розглянутих відомих методів. При цьому кількість кластерів, на яку розбивав вхідну вибірку розроблений метод, була правильною для всіх вибірок.

Висновки

У статті вирішено актуальне завдання автоматизації кластеризації на даних на основі використання дерев розв'язків.

Наукова новизна роботи полягає в тому, що розроблено метод кластерного аналізу, який заснований на побудові дерев розв'язків, що дозволяє виконувати розбиття на кластери шляхом введення рівномірно розподілених точок простору пошуку та скорочує вимоги до обчислювальних ресурсів при виконанні кластерного аналізу. Крім того, при використанні запропонованого методу немає необхідності задання інформації про кількість кластерів та їх форму, що суттєво розширює можливість його застосування на практиці.

Література

1. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP : [уч. пос.] / Барсегян А.А. – СПб. : BHV, 2007. – 384 с.
2. Брянцев И.Н. Data Mining. Теория и практика / Брянцев И.Н. – М. : БДЦ-Пресс, 2006. – 208 с.
3. Rokach L. Data Mining with Decision Trees. Theory and Applications / L. Rokach, O. Maimon. – London : World Scientific Publishing Co, 2008. – 264 p.

4. A Comparison of Decision Tree Ensemble Creation Techniques / R. Banfield, L.O. Hall, K.W. Bowyer, W.P. Kegelmeyer // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2007. – № 29 (1). – P. 173-180.
5. Liu X. A decision tree solution considering the decision maker's attitude / X. Liu, Q. Da // Fuzzy Sets and Systems. – 2005. – № 152 (3). – P. 437-454.
6. Su J. A Fast Decision Tree Learning Algorithm / J. Su, H. Zhang // AAAI: Proceedings of the National Conference on Artificial Intelligence (16 – 20 July 2006). – Boston, USA : AAAI Press, 2006. – P. 31-36.
7. Distributed Decision-Tree Induction in Peer-to-Peer Systems / K. Bhaduri, R. Wolff, C. Giannella, H. Kargupta // Statistical Analysis and Data Mining. – 2008. – № 1 (2). – P. 85-103.

Literatura

1. Barsegjan A. A. Tehnologii analiza dannyh: Data Mining, Visual Mining, Text Mining, OLAP: Uch. pos. Pb: BHV. 2007. 384 s.
2. Brjancev I. N. Data Mining. Teorija i praktika. M.: BDC-Press. 2006. 208 s.
3. Rokach L. Data Mining with Decision Trees. Theory and Applications. London: World Scientific Publishing Co. 2008. 264 p.
4. Banfield R, Hall L.O, Bowyer K.W., Kegelmeyer W.P IEEE Transactions on Pattern Analysis and Machine Intelligence. № 29 (1). 2007. P. 173-180.
5. Liu X. Fuzzy Sets and Systems. № 152 (3). 2005. P. 437-454.
6. Su J. AAAI: Proceedings of the National Conference on Artificial Intelligence (16–20 July 2006). Boston. USA: AAAI Press. 2006. P. 31-36.
7. Bhaduri K., Wolff R., Giannella C., Kargupta H. Statistical Analysis and Data Mining. № 1 (2). 2008. P. 85-103.

A.A. Oliinyk, Ye.A. Gofman, S.A. Subbotin

Clustering Method Based on Decision Trees

The problem of cluster analysis is rather actual while solving the problems of technical diagnostics, pattern recognition and forecasting, which is splitting some data on the set of clusters that represent a compact area in the feature space. There are various methods of cluster analysis [1], [2], the main disadvantage of them is need to pre-task input configurable parameters (e.g. number of clusters to be selected). This complicates their usage and in the processing of data in real situations where there is no sufficient information about an object, process or system.

So it is important to develop new clustering methods, which is free from these disadvantages, and provide the necessary accuracy of the solutions. The main criterion that must be satisfied the methods used for solving this problem is the possibility of separating the space of instances in the area with similar characteristics. These methods include search based on decision trees, which due to its structure perform partitioning of the space of solutions to the field as a function of the input variables [3-5]. Therefore in this paper, we propose to solve the problem of cluster analysis based on the construction of decision trees.

There are various methods for the identification of decision trees (ID3, CART, CHAID, QUEST, C5.0). However, they do not take into account the peculiarities of the problem of cluster analysis related to the evolution of taxons consisting of objects with the most similar characteristics [3-6].

The purpose of this paper is to develop a method of cluster analysis based on the construction of decision trees, which will allow partitioning into clusters by introducing uniformly distributed points of the search space and reduce the demands on computational resources when performing cluster analysis.

Let sample of objects O, each characterized by a set of attribute values X. Then the problem of cluster analysis lies in the fact that on the basis of values of attributes X, one-hit set of objects O in m clusters (subsets) $C_1, C_2, \ldots, C_m$, so that each object O_i belonged to one and only one subset of the partition, and that the objects belong-lying to the same cluster are similar, while, as objects that belong to different cluster are dissimilar.

To achieve this goal the following tasks are solved:

– review of existing methods of cluster analysis and identification of their strengths and weaknesses;

– the study of fundamental concepts, principles and characteristics of decision trees;

– modification of the method in accordance with the specific problem to be solved;

– a comparison of the developed approach with existing methods of cluster analysis by conducting experiments and analyzing results.

In this paper, the problem of automation for data clustering based on decision trees is solved.

The scientific novelty of the work is that the method of cluster analysis, which is based on building decision trees that allows to split into clusters by introducing uniformly distributed points of search space and reduces the requirements for computational resources when performing cluster analysis, is offered. In addition, when using the proposed method it is not necessary task information about the number of clusters and their shape, which significantly extends the possibility of its application in practice.

Стаття надійшла до редакції 19.12.2011.